Meta-analyse via multiniveau-modellen

J.J. Hox en E.D. de Leeuw
Faculteit Pedagogische en Onderwijskundige Wetenschappen, Universiteit van Amsterdam[1]

ABSTRACT

An important goal in meta analysis is to integrate the outcomes of several studies into one single result. This assumes that all studies are replications that differ by sampling fluctuation only; the studies are then said to be homogeneous. If the study outcomes differ by more than chance fluctuation, they are said to be heterogeneous, and the variation in outcomes should be further analyzed. An elegant method to analyze outcome variation is multilevel regression, because it enables us to include study characteristics as explanatory variables. Such study characteristics may be theoretically interesting variables, but also methodological characteristics to correct for artefacts. Various significance tests are available to assess the importance of study variables and to assess if the results are homogeneous or not.

Trefwoorden: *meta-analyse, multiniveau-analyse, homogeniteitstoets, artefacten*

INLEIDING

Kenmerkend voor meta-analyse is dat men een statistische analyse uitvoert op de uitkomsten van eerdere empirische onderzoeken. Het doel daarbij is het samenvatten van de resultaten van meerdere onafhankelijke studies die zich alle met dezelfde onderzoeksvraag bezighouden. Bijvoorbeeld: "Wat is het effect van een sociale vaardigheidstraining op sociaal-angstige kinderen?" Men verzamelt dan de verslagen van experimenten op dit terrein en vat de uitkomsten statistisch samen tot één 'superuitkomst' (zie ook het artikel van Snijders in dit nummer). Vaak is men echter ook geïnteresseerd in meer gedetailleerde vraagstellingen, zoals "Heeft de lengte van de training effect op de uitkomst?", of, "Heeft de samenstelling van de groep invloed op het effect van de training?". In dat geval codeert en analyseert men niet alleen de uitkomst van de studies, maar ook specifieke studiekenmerken, zoals duur van de training, samenstelling van de experimentele en controle groep, sekse van de trainer, etcetera.

Een zeer algemeen model voor meta-analyse is het random effects model (Hedges & Olkin, 1985, 189 e.v.). Dit model gaat er van uit dat de uitkomsten (effecten) in de verschillende studies niet alleen van elkaar verschillen door toevallige steekproeffluctuaties, maar ook als gevolg van reële verschillen tussen de studies, bijvoorbeeld wat betreft het soort steekproef, de experimentele procedure, of de gebruikte instrumenten. Het doel van de meta-analyse is dan niet zozeer het combineren van alle uitkomsten tot één 'superuitkomst,' maar vooral het analyseren van de spreiding in de uitkomsten. Door de verschillen in de uitkomsten te relateren aan de gecodeerde studiekenmerken kunnen we op het spoor komen van mogelijke moderator-variabelen, en zo een verklaring vinden voor verschillen tussen studies. In het random effects model wordt binnen de spreiding van de geobserveerde uitkomsten onderscheid gemaakt tussen toevallige steekproeffluctuaties rond de populatieparameter en systematische variatie van de populatieparameter zelf tussen de studies. Hedges en Olkin (1985, p. 194) geven formules om de variantie van de uitkomsten op te splitsen in ware parametervariantie en toevallige steekproefvariantie, en om de significantie van de parametervariantie te toetsen. Wanneer de parametervariantie groot is, dan noemt men de uitkomsten *heterogeen*. In dat geval wordt meestal aangeraden de analyse voort te zetten door de studies te verdelen in clusters die onderling verschillen, maar intern *homogeen* zijn. De clustering van de studies kan vooraf gedaan worden op basis van de beschikbare studiekenmerken, bijvoorbeeld de duur van de gegeven training of de samenstelling van de groep, maar ook achteraf door een clusteranalyse op de uitkomsten. De verschillende clusters van studies worden vervolgens apart geanalyseerd met het random effecten model, in de hoop dat de gevormde clusters intern homogeen zijn. Men probeert zo inhoudelijk interessante studiekenmerken op het spoor te komen die de heterogeniteit van de uitkomsten verklaren.

Zoals ook Snijders (dit nummer) aangeeft, is er bij meta-analyse sprake van twee geneste niveaus van analyse-eenheden: personen binnen studies. Wanneer we zouden kunnen beschikken over de ruwe gegevens van alle verschillende studies dan zouden we de meta-analyse kunnen uitvoeren door middel van een 'gewone' multiniveau-analyse. In ons voorbeeld hebben we dan één afhankelijke variabele (een test voor sociale vaardigheid) en één onafhankelijke variabele (de indeling in de experimentele trainingsgroep of de controlegroep) op het individuele niveau, en een lineair regressiemodel voor het verband tussen deze beide. Het algemene multiniveau-regressiemodel stelt dat iedere studie zijn eigen regressiemodel heeft. Multiniveau-analyse wordt vervolgens gebruikt om het gemiddelde en de spreiding van de regressiecoëfficiënten over alle studies tezamen te schatten. Wanneer de spreiding van de regressiecoëfficiënten substantieel is, hebben we heterogene regressiehellingen, en is het zinvol om groepskenmerken (bij meta-analyse studiekenmerken zoals de groepssamenstelling, maar ook bijvoorbeeld de methodologische kwaliteit van de betreffende studie) als verklarende variabelen in te voeren om daarmee de verschillen tussen de studies te verklaren. Het een en ander is dan een multiniveau-analyse die niet wezenlijk anders is dan de in onderwijskundig onderzoek gebruikelijke multiniveau-analyses op leerlingen binnen klassen. (Voor een inleiding in het multiniveau-regressiemodel zie bijv. Bosker en Snijders, 1990, en Hox, 1995; statistische details zijn te vinden in Bryk en Raudenbush, 1992 en Goldstein, 1995).

Bij meta-analyse kunnen we echter doorgaans *niet* beschikken over de ruwe gegevens van de individuen uit de oorspronkelijke studies, maar uitsluitend over de uitkomsten van de uitgevoerde analyses, zoals p-waarden, gemiddelden, standaardafwijkingen, of correlaties. In de klassieke meta-analyse zijn dan ook een groot aantal methoden ontwikkeld om uitkomsten van studies te combineren. Hunter en Schmidt (1990) geven een uitgebreide beschrijving van algoritmen en rekenregels, terwijl Hedges en Olkin (1985) een grondig overzicht geven van de statistische achtergronden. Voor een inleiding zie het artikel van Snijders in dit nummer.

Het is echter wel degelijk mogelijk om een multiniveau-regressieanalyse uit te voeren op de gebruikelijke meta-analyse gegevens. Raudenbush en Bryk (1985; Bryk and Raudenbush, 1992) beschrijven het random effects model als een speciaal geval van het algemene multiniveau-regressiemodel. Omdat de analyse niet wordt uitgevoerd op ruwe data, maar op daarvan afgeleide statistische grootheden, moet de analyse enigszins worden aangepast. Bij het programma HLM wordt daarvoor een speciale versie meegeleverd (VKHLM, Bryk, Raudenbush & Congdon, 1994), terwijl in MLn de benodigde aanpassing door de gebruiker zelf moet worden geprogrammeerd (Lambert & Abrams, 1995; zie ook de appendix bij dit artikel). Het voordeel van het toepassen van het multiniveau-regressiemodel en de daarvoor ontwikkelde software in plaats van de meer gebruikelijke meta-analytische methoden en software ligt in de flexibiliteit van de multiniveau-benadering. In het multiniveau-regressiemodel is het relatief eenvoudig om de gemiddelde uitkomst en de spreiding van de uitkomsten te schatten. Indien men *van tevoren* verklarende variabelen op studieniveau kan definiëren dan kunnen deze studiekenmerken gericht in het model worden opgenomen. Ook voor een post hoc verklaring van heterogene studieuitkomsten kunnen verklarende variabelen in het regressiemodel worden opgenomen. Bij het exploratief invoeren van verklarende variabelen in het regressiemodel is de mogelijkheid van kanskapitalisatie uiteraard altijd aanwezig.

HET MODEL

De verschillende studies die in een meta-analyse worden opgenomen gebruiken dikwijls sterk uiteenlopende instrumenten en statistische toetsen. Om de uitkomsten vergelijkbaar te maken worden daarom de resultaten zo goed mogelijk vertaald in een standaard effectmaat, zoals een correlatie of het gestandaardiseerde verschil tussen twee gemiddelden *d*. Het model voor de studie-uitkomsten is

$$d_j=\delta_j+e_j \qquad (1)$$

waarin d_j de uitkomst is van studie j (j=1,...,J), δ_j de populatiewaarde van deze uitkomst, en e_j de steekproeffout van studie j. Verondersteld wordt dat de steekproeffout e_j een normale verdeling heeft waarvan de variantie σ_j^2 bekend is (op basis van de steekproefgrootte in studie j). Wanneer de steekproefgroottes van de afzonderlijke studies niet te klein zijn, dan is de steekproefverdeling van de uitkomsten in veel gevallen bij benadering normaal met een bekende variantie. Hedges en Olkin (1985, p. 175) suggereren een minimum van 20; Bryk en Raudenbush (1992, p.157) houden het op 30. Deze veronderstelling is overigens ook van toepassing op andere gangbare meta-analyse technieken (vgl. Hedges & Olkin, 1985). Tabel 1 geeft een overzicht van enkele veel gebruikte uitkomstmaten en de variantie van de bijbehorende steekproefverdeling.

Tabel 1. Enige schatters van uitkomstmaten en hun steekproefvariantie.[a)]

uitkomst	gebruikelijke schatter	normaliserende transformatie	steekproefvariantie schatter (na transformatie)
verschil gemiddelden	$g=(\overline{Y}_E-\overline{Y}_C)/s$	onnodig	$(n_E+n_C)/(n_En_C)+g^2/(2(n_E+n_C))$
spreiding (st.afw.)	s	$s^*=\ln(s)+1/(2df)$	$1/(2df)$
correlatie	r	$0.5\ln((1+r)/(1-r))$	$1/(n-3)$
proportie	p	$\ln(p/(1-p))$	$1/(np(1-p))$

a) Naar Bryk & Raudenbush (1992, p. 169). g Is de uitkomstmaat voorgesteld door Glass (1976), s^* is een transformatie van de standaard afwijking s voorgesteld door Bryk en Raudenbush (1992). Dergelijke transformaties worden dikwijls toegepast om de steekproefverdeling te normaliseren. Bij de rapportage van de eindresultaten dient de inverse transformatie te worden toegepast. Voor andere schatters van uitkomstmaten en hun steckproefvariantie zie Hedges & Olkin (1985).

Uit de vergelijking voor model (1) blijkt dat de parameters δ_j (de uitkomsten) verondersteld worden te variëren over de studies. De variantie van de populatiewaarden δ_j (de parametervariantie) wordt verklaard door het model

$$\delta_j=\gamma_0+\gamma_1Z_{1j}+\gamma_2Z_{2j}+...+\gamma_pZ_{pj}+u_j \qquad (2)$$

waarin $Z_1...Z_p$ studiekenmerken zijn, $\gamma_0...\gamma_p$ de regressiecoëfficiënten, en u_j de residuele voorspellingsfout waarvan we aannemen dat die normaal verdeeld is met variantie σ_u^2. Wanneer we (2) substitueren in (1) dan krijgen we het volledige model

$$d_j=\gamma_0+\gamma_1Z_{1j}+\gamma_2Z_{2j}+...+\gamma_pZ_{pj}+u_j+e_j \qquad (3)$$

Wanneer er géén studiekenmerken als predictoren in het model zijn opgenomen, dan wordt model (3)

$$d_j=\gamma_0+u_j+e_j \qquad (4)$$

Model (4), het zogenaamde 'intercept-only' model (cf. Bryk & Raudenbush, 1992), is equivalent aan het random effecten-model voor meta-analyse zoals beschreven door Hedges en Olkin (1985), zij het dat de gebruikelijke multiniveau-software een andere schattingsprocedure gebruikt dan de techniek die Hedges en Olkin beschrijven. (Hedges en Olkin stellen als schattingsmethode Weighted Least Squares voor in plaats van de bij multiniveau-analyse gebruikelijke Maximum Likelihood, maar beide methoden leiden doorgaans tot vergelijkbare schatters.)

In model (4) is de intercept γ_0 de schatting voor de gemiddelde uitkomst over de studies, de variantie van de steekproeffouten e_j oftewel σ_j^2 is de steekproeffluctuatie, en σ_u^2 is de variantie van de residuen u_j ofwel de parametervariantie. Daarmee is σ_u^2 een indicatie voor de mate van heterogeniteit van de uitkomsten, en de statistische toets van de nulhypothese dat σ_u^2 gelijk is aan nul equivalent aan de gebruikelijke homogeniteitstoetsen in de meta-analyse (voor een beschrijving van deze homogeniteitstoetsen zie Hedges & Olkin, 1985, p. 194 en Raudenbush, 1994). De proportie ware parametervariantie wordt gegeven door $\sigma_u^2/(\sigma_u^2+\sigma_j^2)$.

Naast de formele toets kan de proportie ware parametervariantie gebruikt worden als een indicator voor de heterogeniteit van de uitkomsten. Dit is analoog aan de goede gewoonte om bij klassieke regressieanalyse niet alleen te kijken naar statistische significantie, maar ook naar verklaarde variantie. Binnen de meta-analyse wordt aangeraden de homogeniteitstoets voorzichtig te gebruiken. Zo wijzen Hunter en Schmidt (1990) er op dat de gebruikelijke chikwadraat-toets voor de homogeniteit bij het samenvatten van een groot aantal studies een groot onderscheidingsvermogen heeft. Met andere woorden, als de nulhypothese van homogeniteit verworpen wordt, dan wijst dit niet noodzakelijk op sterke heterogeniteit. Omgekeerd hoeft de afwezigheid van significantie niet altijd te betekenen dat de tussen-studies variatie verwaarloosbaar is. Hunter en Schmidt (1990) suggereren als vuistregel dat het zoeken naar interessante moderatorvariabelen pas zin heeft wanneer de proportie ware parametervariantie groter is dan 25 procent.

In het meer algemene model (3) zijn studiekenmerken Z_{pj} opgenomen om de verschillen in de uitkomsten te verklaren. In dit model is σ_u^2 de variantie van de residuen u_j nadat de studiekenmerken verdisconteerd zijn, ofwel de residuele parametervariantie. Ook deze kan op significantie worden getoetst. Wanneer de residuele parametervariantie significant is, dan zijn de ingebrachte studiekenmerken niet in staat alle variaties in de uitkomsten te verklaren. Het verschil tussen de parametervariantie in het intercept-only model (model 4) en de residuele parametervariantie in het model met verklarende studievariabelen kan worden opgevat als de hoeveelheid variantie die door de studievariabelen verklaard wordt.

Het multiniveau-model zoals beschreven in model (3) vertoont sterke overeenkomsten met het algemene lineaire model voor vaste effecten zoals besproken door Hedges en Olkin (1985, hoofdstuk 8). Hedges en Olkin gebruiken een afwijkende (matrix-)notatie, in de hier gebruikte notatie is hun model gelijk aan ons model (4) zonder de modelterm u_j. Het lineaire model voor vaste effecten blijkt daarmee een speciaal geval te zijn van hier beschreven multiniveau-model, waarin verondersteld wordt dat de residuele parametervariantie gelijk is aan nul.

De ervaring leert dat, wanneer er sprake is van heterogeniteit, deze doorgaans niet volledig kan worden wegverklaard door de beschikbare studie-variabelen; de residuele parametervariantie wordt zelden nul. Hunter en Schmidt (1990) stellen dat dit het gevolg is van een groot aantal artefacten, zoals o.a. de correctheid van de statistische assumpties in de oorspronkelijke studies, verschillen in betrouwbaarheid van de gebruikte instrumenten, coderingsproblemen en andere onvolkomenheden bij de meta-analyse zelf. Het is onwaarschijn-

lijk dat voor al deze verschillen verklarende studie-variabelen beschikbaar zijn; het uitgangsmateriaal (de rapporten en andere publikaties) is daarvoor doorgaans te summier. Heterogene resultaten zijn daarom eerder regel dan uitzondering, en het multiniveau-model is daarom in het algemeen te verkiezen boven het lineaire model met vaste effecten.

VOORBEELD EN VERGELIJKING MET KLASSIEKE META-ANALYSE

In deze paragraaf analyseren we een voorbeeld dataset met 'klassieke' meta-analysemethoden zoals geïmplementeerd in het meta-analyse programma van Schwarzer (1989). Dit programma is grotendeels gebaseerd op de methoden besproken door Rosenthal (1984), Hunter & Schmidt (1990) en Hedges & Olkin (1985). De dataset bestaat uit de (gefingeerde) resultaten van twintig studies waarin een experimentele groep en een controlegroep met elkaar vergeleken worden.

Een gebruikelijke effectmaat bij experimenteel onderzoek is het gestandaardiseerde verschil tussen de gemiddelden van de experimentele en de controlegroep, door Hedges en Olkin gedefinieerd als $g=(\bar{Y}_E-\bar{Y}_C)/s$, waarbij s de gepoolde standaarddeviatie is. Omdat g een niet geheel zuivere schatter is van het populatie-effect $\delta=(\mu_E-\mu_C)/\sigma$, geven Hedges en Olkin (1985, p. 81) de voorkeur aan een hiervan afgeleide maat d: $d=(1-3/(4(n_E+n_C)-9))g$. De steekproefvariantie van de effectgrootte d is $\sigma_d^2=(n_E+n_C)/(n_En_C)+\delta^2/[2(n_E+n_C)]$ (Hedges & Olkin, 1985, p. 86). De steekproefvariantie is dus afhankelijk van de onbekende populatiewaarde δ. Doorgaans wordt hiervoor de steekproefwaarde d genomen, waarna σ_d^2 bekend wordt verondersteld. Analoog hieraan is de steekproefvariantie van g gelijk aan $\sigma_g^2=(n_E+n_C)/(n_En_C)+g^2/[2(n_E+n_C)]$. In tabel 2 zijn van de twintig studies zowel de g als de d opgenomen; bij gebruikelijke steekproefgroottes blijken dergelijke correcties gelukkig in de praktijk weinig uit te maken.

Tabel 2. Gesimuleerde voorbeeldgegevens twintig studies

ID	weken	g	d	var(d)	p	n_{exp}	n_{con}	n_{tot}	betr.
1	3	-.268	-.264	.086	.810	23	24	47	.900
2	1	-.235	-.230	.106	.756	18	20	38	.750
3	2	.168	.166	.055	.243	33	41	74	.750
4	4	.176	.173	.084	.279	26	22	48	.900
5	3	.228	.225	.071	.204	29	28	57	.750
6	6	.295	.291	.078	.155	30	23	53	.750
7	7	.312	.309	.051	.093	37	43	80	.900
8	9	.442	.435	.093	.085	35	16	51	.900
9	3	.488	.476	.149	.116	22	10	32	.750
10	6	.628	.617	.095	.030	18	28	46	.750
11	6	.660	.651	.110	.032	44	12	56	.750
12	7	.725	.718	.054	.003	41	38	79	.900
13	9	.751	.740	.081	.009	22	33	55	.750
14	5	.756	.745	.084	.009	25	26	51	.900
15	6	.768	.758	.087	.010	42	17	59	.900
16	5	.938	.922	.103	.005	17	29	46	.900
17	5	.955	.938	.113	.006	14	31	45	.750
18	7	.976	.962	.083	.002	28	26	54	.900
19	9	1.541	1.522	.100	.0001	50	16	66	.900
20	9	1.877	1.844	.141	.00005	31	14	45	.750

De voorbeeld dataset bevat ook één theoretisch interessante verklarende variabele op het studie-niveau: het aantal weken dat de experimentele conditie heeft geduurd. Als de experimentele manipulatie effect heeft, dan mag verwacht worden dat het effect groter is naarmate de experimentele manipulatie langer heeft kunnen werken. Tabel 2 bevat voor elk van de twintig studies de tijdsduur van de experimentele manipulatie (in weken), de beide effectmaten (g en d), de steekproefvariantie van d, de eenzijdige overschrijdingskans van de betreffende studie gebaseerd op de gebruikelijke t-toets voor het verschil tussen twee gemiddelden, de steekproefgroottes van de experimentele en de controlegroep en van de studie als geheel, en de betrouwbaarheid van het gebruikte meetinstrument. Terwille van de leesbaarheid zijn in tabel 2 de studies geordend naar grootte van het gevonden effect.

Binnen de klassieke meta-analyse zijn verschillende benaderingen mogelijk, die vaak aanvullend op elkaar gebruikt worden. Een al heel oude benadering is het combineren van de *p-waarden*. De gecombineerde p-waarde (methode van Stouffer, zie Snijders in dit nummer) is in ons voorbeeld $p<0.001$ ($Z=7.73$). Doorgaans wordt dit geïnterpreteerd als een bewijs dat het gezochte effect bestaat. Strikt genomen hebben we hiermee echter alleen aangetoond dat tenminste een van de betrokken nulhypothesen niet geldt (zie het artikel van Snijders in dit nummer). Meer informatief is het om voor elke studie een effectgrootte te berekenen en deze te combineren tot een enkele uitkomst. Wanneer we daarbij de mogelijkheid open willen houden dat de studies niet alle hetzelfde populatie-effect schatten dan kiezen we bij voorkeur voor het combineren van effectgrootten volgens het random effecten model. Een meta-analyse van de effectmaten d in Tabel 2 volgens dit model schat het gemiddelde effect $\delta=0.58$, met een standaardfout van 0.11 ($p<0.001$). Volgens het random effecten model verschilt het gemiddelde effect dus significant van nul. We moeten echter oppassen, want de homogeniteitstoets leidt tot de conclusie dat de uitkomsten heterogeen zijn ($\chi^2_{19}=48.9$, $p<0.001$).

De parametervariantie wordt geschat als 0.17, en de proportie parametervariantie wordt geschat op 0.65. Dat is aanmerkelijk groter dan de ondergrens van 0.25 die Hunter en Schmidt hanteren voor 'interessante' variantie. Het loont dus de moeite om te proberen de verschillen in de uitkomsten te verklaren. De gebruikelijke methode binnen klassieke meta-analyse volgens het random effects model is om te trachten de studies onder te verdelen in clusters die onderling wel homogeen zijn. Vaak wordt hiervoor een eendimensionale clusteranalyse op de uitkomsten gebruikt. Vervolgens tracht men op basis van de beschikbare studiekenmerken te verklaren waardoor de oorspronkelijke heterogeniteit is ontstaan. In ons voorbeeld levert een clusteranalyse op de effectmaat d drie clusters op, waarvan het eerste bestaat uit de studies 1 en 2, het tweede uit de studies 3 tot en met 18, en het derde uit de studies 19 en 20. Voor een interpretatie dienen we antwoord te geven op de vraag waarin de studies in het ene cluster zich onderscheiden van de studies in het andere cluster. Eén mogelijkheid is de duur van de experimentele ingreep. De gemiddelde duur van de manipulatie is 2 weken in het eerste, 6 weken in het tweede, en 9 weken in het derde cluster. Het lijkt er op dat de duur van de experimentele manipulatie de uitkomsten beïnvloedt.

Omdat we de hypothese hebben, dat de duur van de experimentele behandeling (het aantal weken) van invloed is, kunnen we ook direct een a-priori indeling maken gebaseerd op deze studievariabele. We maken twee clusters: het eerste cluster bestaat uit de negen studies van vijf weken of korter, het tweede cluster uit de elf studies van zes weken of langer. In het eerste cluster wordt het gemiddelde effect geschat op 0.33 (se=0.15, $p=0.01$), en in het tweede op 0.77 (se=0.15, $p<0.001$). In studies die langer duren is het effect kennelijk groter. De hypothese van homogeniteit wordt echter in beide clusters verworpen; de indeling in twee groepen van kortdurende en langer durende studies verklaart niet alle parametervariantie. In het eerste cluster is de geschatte proportie parametervariantie gelijk aan 0.61, in het tweede cluster is deze gelijk aan 0.69. Hoewel we geen formele toets hebben uitgevoerd voor het verschil in de gemiddelde uitkomst tussen beide groepen studies, lijkt de conclusie gerechtvaardigd dat de lengte van het experiment inderdaad een verklarende factor is voor de verschil-

len in de studieuitkomsten. De proportie parametervariantie lijkt echter vooral in het tweede cluster niet te verwaarlozen, de bevinding dat deze niet significant is zou ook het gevolg kunnen zijn van een zwakke statistische toets wegens het kleine aantal studies. Enige voorzichtigheid is op zijn plaats bij de gegeven interpretatie.

Een meta-analyse van de uitkomsten in Tabel 2 volgens het multiniveau 'intercept-only' model levert vrijwel dezelfde resultaten als de hierboven uitgevoerde klassieke meta-analyse volgens het random effecten model. Het gemiddeld effect (de intercept γ_0) wordt geschat als δ=0.58, met een standaardfout van 0.11 (p<0.001). De nulhypothese van homogene uitkomsten wordt verworpen: de parametervariantie wordt echter met 0.14 iets lager geschat als bij het random effecten model (χ^2_{19}=49.8, p<0.001). De proportie ware parametervariantie wordt geschat op 0.61.

De kracht van de multiniveau-benadering toont zich wanneer we de heterogeniteit aan een nader onderzoek onderwerpen. We kunnen het aantal weken dat het experiment heeft geduurd als verklarende variabele in de analyse betrekken, waarbij het bovendien niet langer nodig is deze variabele te dichotomiseren, zoals we hierboven in de klassieke meta-analyse deden. De resultaten staan in tabel 3.

Tabel 3. Resultaten random effecten model en multiniveau regressieanalyses op voorbeeldgegevens tabel 2.[a)]

Model	random effecten	multiniveau nulmodel	multiniveau met predictor weken
delta/intercept	.58 (.11)	.58 (.11)	-.22 (.20)
weken behandeling			.14 (.03)
parameter variantie σ_u^2	.17	.14	.04
sig. via χ^2 toets[a]	p<.001	p<.001	p=.09

a) Tussen haakjes de standaardfout. Voor de bepaling van de significantie van de variantiecomponenten is deze *niet* gebruikt. De significantie van de parametervariantie is bepaald met een heterogeniteitstoets overeenkomend met de in de meta-analyse gebruikelijke Q-toets van Hedges & Olkin (1985). Zie de bespreking in de discussie.

Wanneer de tijdsduur van het experiment in het model wordt opgenomen (zie model 3), dan blijkt dat deze variabele een significant effect heeft. De bijbehorende regressiecoëfficiënt heeft een waarde van γ_1=0.14 (p=<0.001), hetgeen inhoudt dat voor elke week dat het experiment duurt het effect gemiddeld met 0.14 toeneemt. In dit model is de intercept γ_0=-0.22 (p=0.20). Het intercept verschilt dus niet significant van nul. Inhoudelijk is dat goed te begrijpen. In een regressievergelijking is de intercept immers de geschatte waarde van het criterium wanneer alle predictoren gelijk zijn aan nul. In ons geval wordt geschat dat als de experimentele manipulatie nul weken heeft geduurd, de experimentele en de controlegroep niet significant van elkaar verschillen. De residuele parametervariantie wordt geschat op 0.04 (p=0.09). Van de oorspronkelijke parametervariantie van 0.14 wordt dus 71% verklaard door de duur van de experimentele manipulatie. De proportie residuele parametervariantie is 0.29; volgens de maatstaven van Hunter en Schmidt dus qua grootte niet geheel verwaarloosbaar.

Aangezien het effect groter wordt naarmate het experiment langer duurt, is het niet zinvol om alleen een gemiddelde effectgrootte te rapporteren. Eén mogelijkheid is om de geschatte effectgrootte te vermelden voor experimenten met een verschillende tijdsduur, of om te bepalen welke tijdsduur minimaal nodig is om een significant effect te bereiken. Het een en ander is eenvoudig uit te rekenen.

Wanneer we de conclusies van de klassieke meta-analyse en de multiniveau-benadering vergelijken, blijkt dat de resultaten in dezelfde richting wijzen. Dat is niet geheel verrassend, conceptueel lijken meta-analyse volgens het random effecten model en de multiniveau-benadering erg op elkaar. De multiniveau-benadering maakt het echter veel eenvoudiger om verklarende variabelen in de analyse op te nemen, en om te toetsen of na het opnemen van de verklarende variabelen nog significante tussen-studies variantie over is. In de volgende paragraaf wordt van de eerste mogelijkheid gebruik gemaakt om de meta-analyse te corrigeren voor twee mogelijke artefacten.

CORRECTIES VOOR ARTEFACTEN

In de meta-analyse wordt, met name door Hunter en Schmidt, dikwijls gepleit voor het corrigeren van effectmaten voor een aantal artefacten (Hunter en Schmidt, 1990, 1994). Een gebruikelijke correctie is de correctie voor attenuatie als gevolg van de onbetrouwbaarheid van het meetinstrument. Hiertoe wordt de effectmaat eenvoudig gedeeld door de wortel uit de betrouwbaarheidscoefficient, waarna de verdere analyse op de gebruikelijke manier wordt uitgevoerd. Deze correctie is gelijk aan de in de psychometrie bekende correctie voor attenuatie. Hunter en Schmidt (1990) beschrijven nog een groot aantal andere correcties. Het probleem met deze benadering is dat de 'ware' effectgrootte die op deze manier geschat wordt vrijwel altijd groter is dan de geobserveerde effectgrootte. Wanneer bijvoorbeeld de oorspronkelijke meetinstrumenten een lage betrouwbaarheid hebben dan leidt de correctie voor onbetrouwbaarheid dus tot zeer grote effectgroottes. Omdat deze grote effecten niet daadwerkelijk geobserveerd zijn, zijn de door Hunter en Schmidt voorgestelde correcties dan ook enigszins omstreden. Schwarzer (1989, p. 56), die de correctie voor onbetrouwbaarheid standaard in zijn programma's heeft ingebouwd, raadt dan ook aan de gecorrigeerde resultaten alleen te rapporteren samen met de ongecorrigeerde.

Een nadeel van het vooraf corrigeren van de effectmaten voor allerlei artefacten, zoals onbetrouwbaarheid van de meetinstrumenten, is dat dergelijke correcties vrij grof zijn, en dat de gegevens a-priori worden aangepast. Wanneer de gebruikte, gerapporteerde betrouwbaarheid onnauwkeurig is, dan is de uitgevoerde correctie onjuist. Zo is de meest gebruikelijke maat voor de betrouwbaarheid, coëfficiënt alfa, een onderschatting van de betrouwbaarheid, zodat de voorgestelde correctie bij deze maat doorgaans een overcorrectie zal zijn. De correcties hebben bovendien invloed op de geschatte standaardfouten, zodat het niet inzichtelijk is wat het cumulatief effect van een aantal correcties is.

Een andere mogelijkheid om voor artefacten te 'corrigeren' is de betrouwbaarheid van de gebruikte meetinstrumenten als predictor in het multiniveau-model op te nemen, in het vaste of ook in het random-gedeelte van het model. Dit is niet optimaal, daar de theoretisch juiste correctie voor onbetrouwbaarheid formeel een multiplicatief model vereist in plaats van het additieve multiniveau-model. Het voordeel van deze werkwijze is echter dat het effect van de onbetrouwbaarheid van de metingen aan de hand van de voorliggende gegevens geschat wordt, zodat het probleem van de overcorrectie niet aan de orde is. Ook andere artefacten kunnen direct gecodeerd worden en als predictor in het model worden opgenomen.

Een specifieke variatie op het invoeren van artefact-variabelen als predictor is het onderzoeken van het effect van de totale steekproefgrootte. In de klassieke meta-analyse raden bijvoorbeeld Light en Pillemer (1984, zie ook Light et al., 1994) aan om de effectmaten in een plot af te zetten tegen de bijbehorende steekproefgroottes. Wanneer de verzameling studies

'well-behaved' is, dan hoort deze plot de vorm van een horizontaal liggende trechter te hebben: studies met een kleinere steekproefgrootte zijn meer variabel, maar schatten wel dezelfde populatieparameter. Wanneer vooral bij de kleinere studies veel onverwacht grote effecten gevonden worden, dan is dit een aanwijzing dat er wellicht selectief gepubliceerd is, en dat er ergens een denkbeeldige bureaulade met ongepubliceerde niet significante studies is. Dit staat bekend als het 'file-drawer' probleem. Behalve door inspectie van een trechterplot, kan deze hypothese ook formeel onderzocht worden door de totale steekproefgrootte als predictor op te nemen in een multiniveau-analyse.

We illustreren de werkwijze bij correctie voor onbetrouwbaarheid en steekproefgrootte weer aan de hand van de voorbeeldgegevens. Tabel 2 bevat de benodigde gegevens wat betreft de studiegrootte en de betrouwbaarheid van het gebruikte instrument.

Wanneer we volgens de methoden van de klassieke meta-analyse te werk gaan dan komen we tot de volgende resultaten. Na correctie voor onbetrouwbaarheid (standaard correctie voor attenuatie, zie Schwarzer, 1989) wordt de gecombineerde effectgrootte geschat op 0.64 in plaats van de ongecorrigeerde waarde van 0.58. De parametervariantie wordt geschat op 0.23 in plaats van de ongecorrigeerde waarde van 0.17. De correctie voor onbetrouwbaarheid leidt er dus toe dat de gecombineerde effectgrootte iets groter wordt, en de parametervariantie eveneens toeneemt. Het effect van de steekproefgrootte is in de klassieke meta-analyse moeilijk te onderzoeken anders dan door een trechterplot. Een trechterplot voor de gegevens uit Tabel 2 (hier niet opgenomen) laat op het eerste gezicht geen duidelijke trend zien.

Wanneer we een multiniveau-analyse toepassen en de verschillende studiekenmerken elk apart opnemen als predictorvariabelen dan krijgen we de resultaten in tabel 4.

Tabel 4. Resultaten multiniveau regressieanalyses op voorbeeldgegevens tabel 3.[a)]

Model:	(1)	(2)	(3)	(4)	(5)
Predictor:	intercept only	(1) plus N_{tot}	(1) plus betrouwb.	(1) plus weken	(1) plus alle predictoren
intercept	.58 (.11)	.50 (.50)	.15 (1.23)	-.22 (.20)	.42 (.92)
N_{tot}		.001 (.01)			-.004 (.01)
betrouwb.			.51 (1.47)		-.51 (1.18)
weken behandeling				.14 (.03)	.15 (.04)
parameter variantie σ_u^2	.14	.16	.16	.04	.05
sig. via χ^2 toets[a]	$p<.001$	$p<.001$	$p<.001$	$p=.09$	$p=.07$

a) Tussen haakjes de standaardfout. Voor de bepaling van de significantie van de variantiecomponenten is deze *niet* gebruikt. De significantie van de parametervariantie is bepaald met een heterogeniteitstoets overeenkomend met de in de meta-analyse gebruikelijke Q-toets van Hedges & Olkin (1985). Zie de bespreking in de discussie.

In Tabel 4 is het 'intercept only' model het basismodel dat de gemiddelde effectgrootte schat als 0.58 en de tussen-studies variantie als 0.14. Model 1 bevat de correctie voor de steekproefgrootte, model 2 de correctie voor de onbetrouwbaarheid, en model 3 het effect van de duur van de experimentele conditie als predictor. In model 4 zijn alle drie predictoren simultaan opgenomen. Zowel uit de afzonderlijke als uit de simultane analyses blijkt dat alléén de predictor 'duur van de experimentele behandeling' een significant effect heeft. Geconcludeerd kan worden dat verschillen in meetbetrouwbaarheid en in steekproefgrootte geen ernstige bedreiging van de validiteit van onze conclusies vormen. Het ontbreken van een significant verband tussen de studiegrootte en de effectgrootte geeft aan dat publikatie-selectie (file-drawer problem) vermoedelijk geen ernstige bedreiging is voor de validiteit van onze conclusie over het effect van de experimentele ingreep.

Het laatste model (model 4), waarin alle predictoren simultaan zijn opgenomen, is om andere redenen nog instructief. De schatting van de regressiecoëfficiënt voor de betrouwbaarheid (weliswaar niet significant) is daar negatief. Dat zou betekenen dat studies met een betrouwbaarder meetinstrument juist een kleiner effect rapporteren. Dat is wonderlijk, en ook tegengesteld aan de regressiecoëfficiënt in het model waar de betrouwbaarheid als enige predictor optreedt. Het is dan ook een suppressor-effect, veroorzaakt door matig positieve correlaties (van 0.25 tot 0.33) tussen de drie predictoren. Model (2), waarin alleen betrouwbaarheid als predictor is opgenomen, impliceert dat als de betrouwbaarheid stijgt van 0.75 naar 0.90, de effectgrootte met (0.51)x(0.15)=0.08 toeneemt. Dit lijkt aardig op de correctie van 0.06 zoals bepaald volgens de conventionele methode (zie hierboven). De grootte van de standaardfouten voor de regressiecoëfficiënten van de artefact-variabelen in tabel 4 (deze zijn in vergelijking met de grootte van de regressiecoëfficiënten zelf uitermate groot) maakt echter duidelijk dat de betreffende correctie kennelijk overbodig is, zo niet misleidend.

SAMENVATTING EN DISCUSSIE

Het gebruik van multiniveaumethoden bij meta-analyse heeft het grote voordeel dat het gemakkelijk is studiekenmerken in de analyse op te nemen als mogelijke verklarende variabelen voor verschillen in de uitkomsten. Deze studiekenmerken kunnen bedoeld zijn als een theoretisch interessante verklaring voor de geobserveerde verschillen, maar ze kunnen ook gebruikt worden om voor potentiële artefacten te corrigeren. Door significantietoetsen op de individuele regressiecoëfficiënten uit te voeren, en door te toetsen of na het opnemen van studievariabelen de residuele parametervariantie gelijk is aan nul, kan worden nagegaan welke variabelen mogelijk interessant zijn, en of de opgenomen variabelen alle verschillen verklaren, of slechts een deel.

Multiniveau-regressieanalyse veronderstelt dat de verbanden additief en lineair zijn. Verder wordt verondersteld dat de geanalyseerde effectmaten een normale verdeling volgen (dit geldt overigens ook voor de meeste conventionele schattingsmethoden in de klassieke meta-analyse, zie Hedges en Olkin, 1985). Voorzover aan deze aannamen niet voldaan is, kan soms door een transformatie een oplossing worden gevonden; zowel Hedges en Olkin (1985) als Bryk en Raudenbush (1992) bespreken er enige. Ook wordt verondersteld (opnieuw: dit geldt ook voor de conventionele schattingsmethoden) dat in de samengevatte studies zelf de statistische aannamen opgaan. Wanneer in de samengevatte studies parametrische methoden gebruikt worden, terwijl de gegevens niet normaal verdeeld zijn, dan introduceert dit 'ruis,' in de vorm van onverklaarbare parametervariantie. Onder andere om deze reden stellen Hunter en Schmidt (1990) voor om bij heterogeniteit de parametervariantie pas te analyseren wanneer deze groter is dan een kwart van de totale variantie.

Een interessante uitbreiding van het multiniveau model voor meta-analyse is het opnemen van meerdere uitkomstmaten per studie. Er is dan sprake van een multivariate meta-analyse. Een voorbeeld hiervan is als bij onderwijskundig onderzoek zowel de resultaten op een taaltest als een rekentest als afhankelijke variabelen gebruikt zijn. Doorgaans worden dan twee meta-analyses uitgevoerd, één voor de taaltest en één voor de rekentest. Wanneer het inhoudelijk zinvol is, dan ligt het voor de hand beide uitkomstmaten in een simultane analyse op te

nemen. Raudenbush en Bryk (1985) beschrijven reeds multiniveau modellen voor meerdere uitkomstmaten, waarbij er echter van uitgegaan wordt dat alle uitkomstmaten in *alle* studies beschikbaar zijn. Wanneer dit niet het geval is en er missing data optreden, kunnen meerdere uitkomstmaten nog steeds in één analyse worden opgenomen, maar wordt het model gecompliceerder. We verwijzen hiervoor naar Kalaian & Raudenbush (1996) en Goldstein (1995).

Voor de feitelijke analyse zijn momenteel twee programma's beschikbaar: een speciale versie van HLM (VKHLM, zie Bryk et al., 1994) die met de PC versie van het programma standaard wordt meegeleverd (voor een beschrijving zie de appendix), en MLn (Rasbash & Woodhouse, 1995). Bij MLn moet het model op een speciale manier worden gespecificeerd (voor een beschrijving zie de appendix). Een klein probleem is dat de verschillende programma's niet *precies* hetzelfde doen. VKHLM gebruikt als schattingsmethode Restricted Maximum Likelihood (REML), hetgeen iets andere uitkomsten geeft dan het conventionele random effecten-model, terwijl MLn de onderzoeker de keuze laat tussen Full en Restricted Maximum Likelihood (die in MLn resp. IGLS en RIGLS genoemd worden). De hier gerapporteerde resultaten zijn bij wijze van controle berekend met beide programma's; de verschillen tussen de parameterschattingen blijken minimaal.

Een groot verschil tussen VKHLM en MLn treedt echter op bij de toets op de significantie van de residuele parametervariantie. MLn berekent voor die variantie een standaardfout gebaseerd op de Maximum Likelihood (ML) methode, die gebruikt kan worden voor een Z-toets van die variantie, terwijl VKHLM een chi-kwadraat-toets uitvoert op de OLS-residuen. Bryk en Raudenbush (1992) beargumenteren dat de in HLM en VKHLM ingebouwde chi-kwadraat-toets superieur is aan de toetsingsmethode in MLn (en ook VARCL). Een toets gebaseerd op het delen van een schatter door de bijbehorende standaardfout veronderstelt namelijk dat deze schatter normaal verdeeld is, hetgeen voor varianties niet opgaat, met name niet wanneer deze klein zijn. Bovendien is de ML-toets asymptotisch, dat wil zeggen dat een goede werking bij kleine aantallen studies onzeker is. Een bijkomend voordeel van de chi-kwadraat-toets bij het uitvoeren van een meta-analyse is dat deze bij het intercept-only model identiek is aan de conventionele chi-kwadraat-toets van het conventionele random effecten-model. De hierboven gerapporteerde variantietoetsen zijn gebaseerd op de chi-kwadraat-toets (het is in MLn betrekkelijk eenvoudig om de chi-kwadraat-toets zelf te programmeren).

Op het terrein van de multiniveau-analyse in het algemeen bestaat een toenemende belangstelling voor Bayesiaanse methoden voor de schatting van effecten en standaardfouten. Hierbij wordt gebruik gemaakt van rekenintensieve methoden zoals Markov Chain Monte Carlo methoden (MCMC, een voorbeeld hiervan is de Gibbs sampler). Met name bij meta-analyse hebben MCMC methoden enkele voordelen (DuMouchel, 1990, 1994). Bayesiaanse schatters voor de varianties zijn minder gevoelig voor de problemen die ontstaan wanneer de varianties dicht bij nul liggen en/of geschat moeten worden op basis van een klein aantal studies. Verder is het bij MCMC methoden mogelijk om in het model een functie op te nemen die de kans op publikatie van het resultaat modelleert. In principe levert dit een elegante benadering op om het file-drawer probleem aan te pakken. Toepassing hiervan bij meta-analyse is bepaald nog niet standaard, en gebruikersvriendelijke software hiervoor ontbreekt voorlopig nog. Een algemeen programma voor MCMC modellen is het programma BUGS (Spiegelhalter, 1995), dat voor meta-analyse gebruikt is door o.a. Tweedie et al. (1994).

REFERENTIES

Bosker, R.J. & Snijders, T.A.B. (1990). Statistische aspecten van multiniveau onderzoek. *Tijdschrift voor Onderwijsresearch*, 15, 317-329.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Bryk, A.S., Raudenbush, S.W. & Congdon, R.T. (1994). *HLM 2/3. Hierarchical linear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.

DuMouchel, W.H. (1990). Bayesian meta-analysis. In: D. Berry (ed). *Statistical methodology in pharmaceutical sciences*. New York: Marcel Dekker.

DuMouchel, W.H. (1994). *Hierarchical Bayesian linear models for meta-analysis*. Washington, DC: National Institute of Statistical Sciences.

Glass, G.V. (1976). Primary, secondary and meta analysis of research. *Educational Researcher*, 10, 3-8.

Gleser, L.J. & Olkin, I. (1994). Stochastically dependent effect sizes. In: H. Cooper en L.V. Hedges (eds.). *The handbook of research synthesis*. New York: Russel Sage Foundation.

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hox, J.J. 1995. *Applied multilevel analysis. (Second edition)*. Amsterdam: TT-Publikaties.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.

Hunter, J.E. & Schmidt, F.L. (1994). Correcting for sources of artifact variation. In H. Cooper & L.V. Hedges (eds.). *The handbook of research synthesis*. New York: Russel Sage Foundation.

Kalaian, H.A. & Raudenbush, S.W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, 1, 227-235.

Lambert, P.C. & Abrams, K.R. (1995). Meta-analysis using multilevel models. *Multilevel Modelling Newsletter*, 7, 17-19.

Light, R.J. & Pillemer, D.B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Light, R.D., Singer, J.D. & Willet, J.B. (1994) The visual presentation and interpretation of meta-analyses. In: H. Cooper & L.V. Hedges (eds.). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Rasbash, J. & Woodhouse, G. (1995). *MLn command guide*. London: Multilevel Models Project, University of London.

Raudenbush, S.W. (1994). Random effects models. In: H. Cooper & L.V. Hedges (eds.). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Raudenbush, S.W. & Bryk, A.S. (1985). Empirical Bayes meta analysis. *Journal of Educational Statistics*, 10, 75-98.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

Schwarzer, R. (1989). *Meta-analysis programs. Program manual*. Berlijn: Institüt für Psychologie, Freie Universität Berlin.

Spiegelhalter, D. (1995). Bayesian analysis of complex random effects models. In T.A.B. Snijders, B. Engel, J.C. van Houwelingen, A. Keen, G.J. Stemerdink & M. Verbeek (red). *Toeval zit overal: Programmatuur voor random coëfficiënt modellen*. Groningen: iec ProGamma.

Tweedie, R.L., Scott, D.J., Biggerstaff, B.J. & Mengersen, K.L. (1994). Bayesian meta-analysis, with application to studies of ETS and lung cancer. Fort Collins, CO: Colorado State University, intern rapport.

Ontvangen 12-8-96
Definitief aanvaard 20-12-96

Appendix: Software voor multiniveau meta-analyse

Het eenvoudigste programma voor een multiniveau meta-analyse is het programma VKHLM (voor Variance Known HLM) dat bij het programma HLM (Bryk et al., 1994) wordt meegeleverd. VKHLM verwacht per studie een regel data waarin achtereenvolgens staan: de effectgrootte, de bijbehorende steekproefvariantie, en de verklarende studievariabelen. VKHLM werkt overeenkomstig aan HLM: het model wordt interactief gespecificeerd en VKHLM berekent vervolgens de parameters en standaardfouten, alsmede een groot aantal significantietoetsingen. Wanneer VKHLM wordt gebruikt om een 'intercept-only' model te specificeren wijken de resultaten nauwelijks af van die van het random effecten model in het programma META van Schwarzer (1989), mits er rekening mee gehouden wordt dat META de ingevoerde g's intern transformeert naar d's (de transformatie is eerder toegelicht in de tekst).

Ingewikkelder is het gebruik van het programma MLn. De datastructuur is analoog aan VKHLM: de effectgrootte, de bijbehorende standaardfout (de wortel uit de steekproefvariantie), de constante (VKHLM voegt die automatisch toe) en de verklarende studievaria-

belen. De opzet van de analyse is als volgt. We onderscheiden twee niveaus: niveau 1: de uitkomsten, en niveau 2: de studies. De predictor 'steekproeffout' wordt alleen opgenomen in het random deel op niveau 1 (de uitkomsten) met een coëfficiënt vastgezet op één (met RCONstraint), en niet in het gefixeerde deel. De constante wordt alleen opgenomen in het gefixeerde deel op niveau 1, en niet in het random deel op niveau 1. De constante wordt op het 2e niveau (de studies) normaal in het random deel opgenomen. Verklarende studievariabelen worden alleen opgenomen in het vaste deel. De toets op de variantie van de constante is bij MLn de asymptotische toets. De chi-kwadraat-toets uit Bryk en Raudenbush (1992) heeft als formule: $\chi^2=\Sigma((d_j-\hat{d}_j)/sed_j)^2$, waarbij d_j het waargenomen effect en $\hat{d}_j$ het voorspelde effect is ($d_j-\hat{d}_j$ is daarmee het residu) en sed_j de standaardfout van het effect in studie *j*. Het aantal vrijheidsgraden is J-q-1, waarbij J het aantal studies is en q het aantal geschatte regressiecoëfficiënten. (De MLn code hiervoor is: PRED C20; CALC C20=(('d'-C20)/'sed')^2; SUM C20 to B1; CPRO B1 df. Deze code veronderstelt dat in de spreadsheet kolom C20 ongebruikt is, en dat de kolommen aangeduid met 'd' en 'sed' de d_j's resp. de sed_j's bevatten).

HLM en MLn zijn verkrijgbaar via het interuniversitaire expertisecentrum ProGamma (gamma.post@gamma.rug.nl). De meta-analyseprogramma's van Schwarzer zijn (inclusief uitvoerige documentatie) als freeware beschikbaar via Internet (http://fub46.zedat.fu-berlin.de:8080/~ahahn/meta_e.htm). Spiegelhalter's programma Bugs is eveneens verkrijgbaar via Internet (http://www.mrc-bsu.cam.ac.uk/pub/methodology/bugs). Alle genoemde programma's zijn beschikbaar voor MSDOS, enkele ook voor andere computerplatforms.

1) Faculteit POW, Universiteit van Amsterdam, Wibautstraat 4, 1091 GM Amsterdam, tel. 020-5251530, fax 020-5251200, E-mail HOX@EDUC.UVA.NL